Context-Based Rules for Grammatical Disambiguation in the Tatar Language

Ramil Gataullin[1], Bulat Khakimov[1,2(✉)], Dzhavdet Suleymanov[1,2], and Rinat Gilmullin[1]

[1] Institute of Applied Semiotics, TAS, Kazan, Russia
bulat.khakeem@gmail.com
[2] Kazan (Volga Region) Federal University, Kazan, Russia

Abstract. The paper is dedicated to the problem of grammatical ambiguity in the Tatar National Corpus and describes the methodology and software used for automation of the disambiguation process. Grammatical ambiguity is widely represented in agglutinative languages like Turkic or Finno-Ugric. Disambiguation in the corpus is based on the context-oriented classification of ambiguity types which has been carried out on corpus data in the Tatar language for the first time. In this study the corpus is used as a source for the research and at the same time as a destination for implementing the results. The grammatical ambiguity types are detected automatically using the finite-state morphological analyzer and then classified. In order to build up the grammatically disambiguated subcorpus, a special software module was developed. It searches for ambiguous tokens in the corpus, collects statistical information and allows creating and implementing the formal context-based disambiguation rules for different ambiguity types.

Keywords: Disambiguation · Grammatical homonymy · Context-based rules · Linguistic software · Turkic languages · Corpus linguistics

1 Introduction

The problem of grammatical disambiguation is very relevant for the modern computer and corpus linguistics, especially in relation to such morphologically rich languages as the Turkic group that forms a subfamily of Altaic languages. For example, the Tatar language is spoken in the far part of Eastern Europe (in the Volga region of Russia) and southern parts of Western Siberia. The number of speakers in Russia was 5.31 mln in 2010.

The Tatar National Corpus named "Tugan Tel" [1] was developed by the "Applied semiotics" Research Institute of the Tatarstan Academy of Sciences and the Kazan Federal University in Russia [2]. It's general concept is presented in [3]. For the automated morphological annotation a finite-state morphological analyzer is used [4]. The grammatical tagset for corpus annotation of Tatar word forms is continuously improved and aimed to be more relevant to the Tatar language peculiarities [5]. On the basis of statistical corpus data, the contextual constraints for certain grammatical

N.T. Nguyen et al. (Eds.): ICCCI 2017, Part II, LNAI 10449, pp. 529–537, 2017.
DOI: 10.1007/978-3-319-67077-5_51